mutations, for example from mutagenesis scans or from naturally occurring mutations which affect the function of interest.

Another example is based on proximity. For example, it can be assumed that residues which are close to the active site of an enzyme are more likely to affect enzyme activity and/or specificity than more distant residues and thus, a mutation of a residue near the active site will affect the activity and/or specificity (either positively or negatively) than a mutation further away from the active site. The same proximity argument can be used for other applications: proximity to an epitope, proximity to an area of structural conflict, proximity to a conserved sequence, proximity to a binding site, proximity to a cleft in the protein, proximity to a modification site, etc.

There are a variety of methods available to estimate distance, and any technique known or developed in the art for estimating such distances can be used. For instance, the library can be constrained by distance of α or ß-carbons to the active site of an enzyme. In another embodiment, the constraint can be based on the residues that make contact with the residues of interest (=1st shell) and residues which contact the residues in the 1st shell (= 2nd shell).

In another example, the simple distance function between β carbons of the enzyme and the β carbon of a bound ligand can be used to constrain a library. A linear function can be used where the threshold of acceptable mutations depends on the distance from the bound ligand. However, one can also utilize a variety of other functional relationships between distance and threshold of mutability, *e.g.*, the square of the distance or the square root of the distance.

The physical distances from a known crystal structure of the reference sequence can be used. Alternatively, molecular modeling approaches can be used. For example, the structure of the reference sequence can be predicted based on its homology to a known structure, and then used to calculate distances. Or the entire structure of the reference sequence can be predicted and distances then calculated from the predicted structure. Energy minimization methods can be used.

Another way to generate constraint vectors is through correlation in evolutionary data. It has been observed that the replacement of a residue in a protein or protein family can be correlated with replacements in other positions. See, Lockless & Ranganathan, *Science* **286**:295 (1999); and Gobel, et al., *Proteins* **18**:309 (1994). In such cases it may be advantageous to design the constraint vector such that all correlated residues are mutated simultaneously.

Conservation Indexes can be used as the elements of a constraint vector. In this capacity, one can avoid mutating residues that are highly conserved, or conversely, focus mutations on conserved regions of the protein. Algorithms for calculating Conservation Indexes at each position in a multiple sequence alignment are known in the art (Novere et al. Biophys. Journal v.76 , p. 2329-2345, May 1999).

One of skill will realize that, like a probability matrix, generation of a constraint vector can require quite complex mathematical calculations and therefore an algorithm that determines the vector may be desired or even needed. The development of such an algorithm is within the skill in the art following the teachings herein. Similarly, because of the complex calculations necessary to carry out the algorithm, it can be desirable to generate a computer program and to employ it on a computer to generate the constraint vector. Again, this is within the skill in the art following the teachings herein.

IV. APPLICATION OF THE CONSTRAINT VECTOR TO THE PROBABLILITY MATRIX TO PRODUCE A SUBSTITUTION SCHEME

To determine which positions are to be permuted and which new residues will be tried in those positions, the constraint vector is applied to the probability matrix. This is done to increase the chance of finding improved variants and to decrease the risk of producing mutants with undesired properties, while generating a library of a size which can be effectively screened for a desired property. This application can also determine the degree to which a given change will be represented in the library, or a simpler threshold approach can be used, wherein all changes at a given position which meet the criteria imposed by the constraint vector are equally represented in the library.

An exemplary algorithm is shown in Figure 1. As is graphically represented in Figure 1, the constraint vector can be imagined as being “lowered” onto the probability matrix.

Positions in the probability matrix which are higher than the corresponding value in the constraint vector (i.e., which exceed the threshold imposed by the constraint vector) are candidates for mutagenesis. As the constraint vector is lowered, the number of positions to be mutagenized increases, and the number of new substitutions at each position increases. The degree to which the constraint vector is lowered is thus a determining factor in the size of the library which results. Application of the constraint vector can thus itself be constrained by the desired size of the library; a predetermined library size can be used to determine the degree to which the constraint vector allows the probability matrix to be sampled.

The substitution scheme produced by applying the constraint vector to the probability matrix is itself a useful result. The substitution scheme can be provided and used to create a library. The substitution scheme can be subjected to additional constraints prior to being employed in creating a library. For example, knowledge-based approaches can incorporate information about the activity of the polymer of interest and can be used to focus the substitution scheme to identify residues more likely to result in the desired activity when substituted as well as in identifying residues less likely to result in the desired activity.

One of skill will realize that the application of a constraint vector to a probability matrix can require quite complex mathematical calculations and therefore an algorithm that applies these two algorithms may be desired or required. The development of such an algorithm is within the skill in the art following the teachings herein. Similarly, because of the complex calculations necessary to carry out the application algorithm, it can be desirable to generate a computer program and employ it on a computer to do this. Again, this is within the skill in the art following the teachings herein.

V. CONSTRUCTION OF A LIBRARY

The simplest randomization scheme for polynucleotides encoding proteins is codon-based mutagenesis. In other words, after the amino acid residues to be mutated have been identified, the corresponding codons in the corresponding DNA sequence are randomized to create a DNA library. Procedures to randomize codons are known in the art (Huse et al., Int Rev Immunol. 1993;10(2-3):129-37; Kirkham et al., J Mol Biol. 1999 Jan 22;285(3):909-15). As